UNIVERSIDAD AUTONOMA DE MADRID

ESCUELA POLITECNICA SUPERIOR

Grado en Ingeniería Informática

TRABAJO FIN DE GRADO

Predicción automática del valor del inmueble usando variables macroeconómicas

Marek Emilio Baczewski
Tutor: Ángela Fernández Pascual
Ponente: José Ramón Dorronsoro Ibero

Julio 2020

Predicción automática del valor del inmueble usando variables macroeconómicas

AUTOR: Marek Emilio Baczewski
TUTOR: Ángela Fernández Pascual

Grupo de Aprendizaje Automático
Dpto. Ingeniería Informática
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Julio de 2020

Resumen

El uso de modelos automáticos de valoración para predecir el valor de un inmueble es una práctica cada día más extendida y aceptada tanto por entidades financieras, empresas de servicios de inversión, inmobiliarias, compañías de seguros y demás usuarios, debido a su bajo coste, ahorro de tiempo y objetividad.

Los modelos de valoración automática hacen uso de algoritmos estadísticos junto a una base de datos de inmuebles representativa y actualizada del entorno geográfico en el que se pretende predecir el valor, ya que éste, según el modelo utilizado, se realiza en un lugar y momento concreto. Los datos incluidos dentro del modelo son las características propias del inmueble, así como las de su entorno, siendo la superficie la variable que aporta más valor.

El objetivo principal de este trabajo de investigación es el de incorporar variables macroeconómicas en dichos modelos de valoración automática, haciendo un análisis de su aportación en el valor final del inmueble.

Esta investigación se realiza sobre los datos de tasaciones reales de 2019 dentro de la Comunidad de Madrid. Se propone un modelo base, en este caso un random forest con los datos característicos del inmueble, posteriormente se realiza un modelo mejorado incluyendo las variables macroeconómicas.

El estudio aporta claras evidencias de lo favorable que es la incorporación de las variables macroeconómicas dentro del modelo.

Se denota una mejora sustancial en el desempeño del modelo, basado en un respaldo robusto de las características asociadas a la economía de un país, en el caso específico de España, este resultado es muy favorable dado que una de las finalidades de la valoración automática es la de buscar una mayor precisión para la revisión de valores de inmuebles hipotecados.

Abstract

The use of automatic valuation models to predict the value of a property is an increasingly widespread and accepted practice by financial entities, investment services companies, real estate companies, insurance companies and even other users, due to its low cost, short time procedure and objectivity.

Automatic valuation models use statistical algorithms combined with a representative and updated database of properties in a geographical environment in which the value is to be predicted, since this, according to the model used, is carried out at a specific place and time. The data included in the model are the characteristics of the property as well as those of its environment, being the surface the variable that provides most information.

The main objective of this research work is to incorporate macroeconomic variables in these automatic valuation models, making an analysis of their contribution to the final value of the property.

This research is carried out on the 2019 real appraisal data within the Community of Madrid. A base model is proposed, in this case a Random Forest with the characteristic data of the Property, later an improved model is made including the macroeconomic variables.

The study provides clear evidence of how favorable it is to incorporate macroeconomic variables into the model.

It is denoted, a specific improvement in the performance of the model, considered in a and a robust support of the specific characteristics of the country's economy, in the specific case of Spain, this result is very favorable given that one of the purposes of the Automatic evaluation is the search for greater precision for the revision of the values of the mortgaged properties.

Palabras clave

Modelo de valoración automática (AVM), Random Forest, hipoteca, tasación, variables macroeconómicas.

Keywords

Automated Valuation Model (AVM), Random Forest, mortgage, appraisal, macroeconomic variables

Agradecimientos:

A todos aquellos que sin saberlo me ayudaron.

INDICE DE CONTENIDOS

INDICE DE FIGURAS

INDICE DE TABLAS

1 Introducción

1.1 Motivación

El significado de vivienda según el ámbito desde el que se enfoque puede ser muy variopinto, tiene tantas definiciones válidas que podríamos hacer un TFG sólo de ese tema,

Según Oscar Adrián Dander Sánchez (2012) si consideramos un ámbito estructural y de evolución, podemos ver que, desde tiempos inmemorables, los seres humanos han sentido la necesidad de refugiarse de los peligros de su entorno, llámese condiciones climáticas adversas, huyendo de ser presa de un animal más grande, o simplemente para sentirse en un hogar. En la prehistoria estas viviendas eran cuevas naturales, en las que se refugiaban con su familia o su clan, esto fue evolucionando al igual que muchos otros aspectos de la vida del ser humano de entonces. Pasaron de reutilizar refugios naturales a construir los suyos propios de acuerdo con sus necesidades, partiendo de los materiales que les ofrecía el entorno y basándose en modelos conocidos, por ejemplo, un solo acceso al igual que en una cueva, orientada para aprovechar al máximo la luz diurna, entre otras muchas opciones. En las grandes urbes, la creación de dichas cuevas artificiales ha pasado a ser competencia exclusiva de arquitectos e ingenieros especializados para la parte del diseño, y la construcción por parte de empresas y profesionales específicos.

Desde un punto de vista legal, el derecho a un vivienda digna y adecuada se recoge en la Declaración Universal de los Humanos (artículo 25.1)[1]. En España, la constitución también recoge este derecho vital (artículo 47)[2].

Desde un punto de vista más social y humano, la vivienda es el lugar donde ocurren la gran mayoría de actividades básicas de una vida cotidiana, y es el lugar donde se regresa al final de una larga jornada de trabajo o estudio.

[1] https://www.un.org/es/universal-declaration-human-rights/

[2] https://app.congreso.es/consti/constitucion/indice/titulos/articulos.jsp?ini=47&tipo=2

Es por esto por lo que la vivienda no es solo un conjunto de paredes estructuradas sistemáticamente, por el contrario, existen miles de tipologías y formas, las cuales se adaptan a las necesidades de cada usuario.

Como bien expone la empresa especializada en tasaciones, Tinsa (15 febrero, 2016), estas distintas tipologías de inmuebles y en especial la localización de las mismas, así como la superficie y la funcionalidad del inmueble, son las que enfatizan o deprecian ese valor y es de interés fundamental y económico el conocer el valor estimado de un inmueble, dado que tener conocimiento del valor real es prácticamente imposible debido a la gran cantidad de variables que aportan al valor final del inmueble, y muchas de estas son completamente subjetivas, tales como el estado de conservación entre otras.

Tinsa (15 febrero, 2016) prosigue evidenciando la existencia de muchas maneras de dar valor a una vivienda, objetivamente cualquier persona puede dar un valor a una vivienda y en función de sus conocimientos sobre el mercado inmobiliario, este valor estará más o menos ajustado al valor real de la vivienda. Un profesional de la tasación de inmuebles dará un resultado más acertado, así como un modelo estadístico también es capaz de dar una predicción del valor basada en variables tangibles de los inmuebles utilizados para entrenar dicho modelo (superficie, altura, estado de conservación).

Marketing y Tecnología (5 de septiembre, 2019) relata que el uso de modelos estadísticos, denominados modelos de valoración automática (automated valuation models, AVM) se están utilizando cada vez más hoy en día por empresas del sector inmobiliario para realizar estas valoraciones, los cuales combinan el análisis basado en la descomposición de un bien inmueble en sus características más importantes y la contribución que tienen dentro del valor agregado, con un análisis del valor del suelo. Asimismo, es bien sabido que una normalización de los datos y una geocodificación de todas las muestras utilizadas para entrenar el modelo dará lugar a mejores resultados. Actualmente, dentro del sector inmobiliario los modelos de valoración automática utilizados sólo contemplan el uso de las variables directamente relacionadas con las características de la vivienda y con su entorno más inmediato, obteniendo con estos modelos resultados aceptables. La hipótesis que se quiere comprobar en este TFG, dado que una vivienda no es un ente aislado de la economía, y que su valor evoluciona con la misma, es la inclusión de variables macroeconómicas para ajustar el valor predicho con el real.

1.2 Objetivos

El objetivo fundamental de este estudio es evidenciar lo favorable de incluir variables macroeconómicas en el entrenamiento de modelos de valoración automática, obteniendo un modelo más sólido y acorde a la economía de una determinada región geográfica.

Para ello se plantean los siguientes hitos durante el desarrollo:

1. Lectura del estado del arte de modelos de valoración automática para inmuebles.
2. Réplica de un modelo de valoración automática sobre un conjunto de datos de inmuebles para predecir su valor.
3. Búsqueda y tratamiento de datos macroeconómicos de una determinada región.
4. Entrenamiento de un nuevo modelo equivalente a (3.) añadiendo variables macroeconómicas.
5. Evaluación y comparación de ambos modelos.

1.3 Organización de la memoria

Este trabajo se organiza de la siguiente forma:

La parte inicial, incluye un resumen autocontenido, una lista de palabras clave, ambos facilitados en español e inglés, culminando con los índices, de contenido, figuras y tablas.

La parte central se encuentra dividida en 6 capítulos: en el primer capítulo se versa una breve introducción autocontenida, seguida de una lectura actualizada sobre el estado del arte incluyendo una crítica al mismo, posteriormente se explica el análisis del problema y diseño de la solución, a continuación, el desarrollo y la implementación, como penúltimo capítulo la integración de pruebas y resultados, finalizando con los resultados obtenidos, su análisis y las conclusiones extraídas.

La sección final incluye las referencias, el glosario y los anexos.

2 Estado del arte

2.1 Contexto actual

En la actualidad son muchas las empresas del sector inmobiliario, o relacionadas con este sector que emplean un gran esfuerzo en el desarrollo de sistemas de inteligencia artificial, en concreto de aprendizaje automático, apoyándose en una base de datos fiable de muestras de inmuebles de tasaciones reales para la elaboración de un modelo robusto de valoración automática (Automated Valuation Model, AVM). Estos sistemas se basan en las normativas dictadas en su mayoría por el Banco de España y por la Asociación Española de Análisis de Valor.

Para entender el origen de una valoración automática, en primer lugar, se tiene que estudiar su origen y evolución.

El origen de los AVMs surge de la necesidad tanto del sector público como del privado, de conocer el estado financiero o la evaluación de la solvencia de entidades financieras, empresas, fondos inmobiliarios, inversores y particulares.

Si realizamos el análisis con mayor profundidad de un AVM, vemos que su origen reside en una tasación.

La tasación, en este caso referida a una tasación de índole hipotecaria y/o de determinación de patrimonio y realizada de forma presencial, se define como una valoración de un inmueble, en este estudio una vivienda, calculado por distintas metodologías, para que sea objeto de garantía para el prestamista, en su mayoría de casos entidades financieras, el cual asume el riesgo del valor tasado.

Dentro de los ámbitos de aplicación del valor de un bien inmueble, podemos encontrar un amplio abanico de situaciones.

Para entidades financieras las más comunes son: establecer un préstamo hipotecario o en su defecto una garantía hipotecaria, para conocer la cobertura de su cartera de viviendas.

Para aseguradoras, reaseguradoras y otras empresas que trabajen acordando un seguro al total o parcial del inmueble, establecer un valor recuperable de una vivienda asegurada.

Para instituciones de inversión colectiva inmobiliaria, conocer el patrimonio inmobiliario (con ello determinar la solvencia), para fondos de pensiones igualmente, conocer el patrimonio inmobiliario del que disponen.

Existen varias metodologías en cuanto a tasación de inmuebles se refiere, todas buscan objetivamente obtener el valor real de un inmueble, sabiendo de antemano que ese valor es distinto cuando se refiere a valor de mercado (es el precio al que podría venderse el inmueble mediante un contrato privado en el supuesto que el bien se hubiera ofrecido públicamente en el mercado), valor hipotecario (es el que se realiza para contratar un préstamo hipotecario, dado que es el valor sostenible en el tiempo), o valor de reemplazamiento (por una parte el valor bruto es la suma de las inversiones necesarias para construir otro inmueble de las mismas características, pero utilizando tecnología y materiales actuales, por otra parte el resultado de deducir la depreciación física y funcional del valor bruto, es el neto.)

La obtención de cada valor antes mencionado está asociado al método de valoración. En el caso del valor de reemplazamiento, hacemos uso del método de coste, aplicable a toda clase de edificios y elementos de edificios, sin importar el estado actual del inmueble (en proyecto, en construcción, en rehabilitación, o terminados). Los componentes que se incluyen en el cálculo son: el valor del suelo, el coste de construcción, los OGN (otros gastos necesarios, entre los cuales están licencias, tasas o aranceles), gastos financieros imputables y el beneficio para el promotor. Es natural entender que esta metodología se aplica cuando el activo valorado no dispone de un mercado de comparables suficiente que permita el método de comparación. La definición del cálculo la podemos encontrar en la sección segunda, Método del Coste, articulo 18 y 19.

El cálculo por el método de actualización, siempre que se cumplan los requisitos establecidos por la normativa (SECCIÓN 4.ª MÉTODO DE ACTUALIZACIÓN DE RENTAS, artículo 25) será aplicable a todos aquellos inmuebles susceptibles de producir rentas salvo las opciones de compra. Cabe resaltar que uno de los requisitos fundamentales es la existencia de un mercado de alquileres representativo de la zona y disponer de esos datos para realizar los cálculos, por lo que el valor final vendrá determinado por el valor presente de todas las rentas futuras imputables al inmueble, es decir el valor atribuible viene en función de las rentas que produce o puede producir, los componentes que intervienen en el procedimiento del cálculo, estimación de los flujos de caja, estimación

del valor de reversión, elección del tipo de actualización. Los cálculos se obtienen de aplicar las fórmulas descritas en la definición del cálculo de la normativa actual (SECCIÓN 4.ª MÉTODO DE ACTUALIZACIÓN DE RENTAS, artículo 26, 27, 28)

El método residual, es aplicable para aquellas viviendas edificadas o en estado de construcción. Está basado sobre el valor residual y del mayor y mejor uso, es decir consiste en calcular el valor del inmueble con la construcción finalizada y restarle los gastos en los que hay que incurrir para que el inmueble llegue a ese estado. Se subdivide en 2 tipos, uno dinámico el cual se podrá aplicar a terrenos urbanizables con o sin edificación, y a edificios en proyecto, construcción o rehabilitación, sin tener en cuenta si están o no paralizado; el otro es el estático, el cual solo se podrá aplicar a solares e inmuebles en rehabilitación, con el condicionante del inicio de obra no superior a un año, el cálculo se define en la normativa actual, (SECCIÓN 4.ª MÉTODO Residual, artículo 36).

El método de comparación es el más utilizado en la actualidad y a su vez el más fiable gracias a la estandarización a la que se ve sometido por la normativa estatal Esta metodología se basa en buscar inmuebles lo más homogéneos posibles al que se quiere valorar, para ello hace falta contar con un mercado representativo de los inmuebles colindantes, que tomarán el papel de comparables. Una vez habiendo observado que el mercado existe, el siguiente paso natural es obtener la información de las transacciones llevadas a cabo en dicha zona para posteriormente hacer un cálculo regresivo y obtener según las características del inmueble los coeficientes a aplicar para la valoración y para la homogeneización. Cabe resaltar que la cantidad de transacciones para realizar el cálculo regresivo debe también ser representativa y a su vez extenderse por lo menos 2 años hacia atrás. Una vez obtenidos los coeficientes, se procede a elegir mínimo 6 inmuebles, por lo general se eligen los que sean geográficamente más cercanos y con ciertas características parecidas de forma general, no necesariamente del todo homogéneas. En segundo lugar, se aplican los coeficientes obtenidos en el cálculo inicial para homogeneizar los inmuebles. Una vez que los inmuebles sean homogéneos, se procede a aplicar los coeficientes para valorar, y además de ello se le añade el valor particular de los elementos de la edificación que tengan algún carácter histórico o artístico.

Si se extrapola esta práctica presencial a la valoración de una cartera de inmuebles, se deduce que es inviable su realización en un tramo de tiempo aceptable, dado que realizar

un número elevado de tasaciones se extiende en el tiempo, sobre todo para aquellas extensas carteras de entidades financieras y fondos de inversión.

Es por esto que, partiendo de la definición de tasación, se puede deducir la definición de un AVM, son todos los programas informáticos basados en algoritmos matemáticos, y estadísticos que permiten obtener con un cierto grado de confianza, el valor de mercado de un inmueble partiendo de un conjunto de datos, de los cuales se obtienen unos parámetros ajustados al mercado que representan. Estos parámetros van de acuerdo al marco legal español, que tiene un enfoque a la valoración masiva de carteras en su valor conjunto, o a una valoración regulada según lo establecido por el Banco de España (Circular 4/2017).

2.1.1 Modelos actuales

En este apartado se van a especificar las metodologías más aceptadas para la estimación del valor final de un inmueble y su correcta aplicación. En primer lugar, nótese que todo procedimiento de valoración, al basarse en métodos estadísticos, siempre lleva implícito cierta incertidumbre que debe ser controlada. Es fundamental tener presente que es harto complejo controlar la incertidumbre contemplando el total del conjunto del universo de valores, por lo que en absoluto la calidad del modelo se ve mermada en caso de utilizar parámetros estadísticamente aceptables. Tal y como expresa la AEV (2 de julio 2019) los mecanismos de control de dicha incertidumbre más usuales en este contexto vienen dados en primera instancia por excluir aquel colectivo cuyos valores se hallen muy alejados del valor del conjunto total de inmuebles (outliers). Otro mecanismo es la selección de las variables explicativas más relevantes, descartando aquellas que estén correlacionadas entre sí o aporten una información similar y su influencia en el valor del inmueble sea trivial. Estos procedimientos resultan imprescindibles para obtener unos valores homogéneos que esclarezcan aquellos comportamientos mayoritarios y sean representativos de la realidad del fenómeno que tiene como objetivo inferir. Según Solvia (27 de marzo 2018) las variables que comprenden gran parte de la explicación de los comportamientos de los mercados inmobiliarios basándose en estudios disponibles son el ámbito territorial o la localización del objeto de estudio, la tipología unifamiliar o plurifamiliar, la superficie en metros cuadrados, la antigüedad y variables socioeconómicas o constructivas tales como nivel de renta o calidad de edificación. En el contexto de las aseguradoras toda incorporación de variables distintas a las anteriores deberá ser debidamente justificada con su influencia en el valor del inmueble.

En cuanto a los modelos que se utilizan, los hay de diversa índole, desde un análisis de regresión, pasando por metodologías de comparación de muestras directamente, hasta redes neuronales y el método hedónico. Todos estos métodos están explicados brevemente a continuación.

2.1.1.1 Análisis de regresión

AZNAR-BELLVER, J., et al (2012) en su publicación explica que el análisis de regresión pretende esclarecer la relación, en caso de que exista, entre las variables explicativas (características) y la variable dependiente (valor del inmueble). Un modelo de regresión aproxima una función (valor estimado) a la función real (valor real) minimizando el error entre dichas funciones. Para estimar la función real, se necesitan valores de entrada y salida con los que se pueda parametrizar dicha función y obtener un resultado cercano a la función real. Si las relaciones entre las variables son lineales, lo más adecuado es emplear la regresión lineal, en caso de que sean no lineales lo óptimo es el planteamiento de alguno de los modelos que se expondrán a continuación.

2.1.1.2 Método de comparables

La AEV (2 de julio 2019) especifica concretamente que lo siguiente: el método de comparables tiene como fin simular la actuación de los profesionales en una valoración de una tasación, para ello, no requieren de la aportación de valores de tasación previos ya que estiman el valor del inmueble objetivo mediante la comparación y homogeneización de inmuebles próximos. La selección de la muestra de inmuebles viene dada por el cumplimiento de una serie de requisitos tales como el radio geográfico de búsqueda, tipología del inmueble, superficie, antigüedad y otorgar preferencia a aquellos comparables más recientes. Una vez obtenida las muestras, un método de cálculo es el que se especifica en la siguiente fórmula:

$$Valor\ inicial\ del\ inmueble = \mu \pm \sigma$$

$$\text{Donde } \mu = \frac{\sum_{i=1}^{n} x_i}{n} ; \sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n}}$$

donde μ es el valor medio unitario de todos los comparables seleccionados y σ la desviación típica de dichos valores medios unitarios. Se descartan aquellos inmuebles situados fuera del intervalo de confianza y se recalcula la siguiente expresión:

$$\mu_2 = \frac{\sum_{j=1}^{n} x_j}{n}$$

$$Valor\ final\ del\ inmueble = \mu_2$$

El ejemplo anterior es una aproximación general de lo que sería el cálculo del valor final, sin embargo, no es el único. El método siempre está abierto a modificaciones en la selección de muestras para calcular el valor o realizar diversos procedimientos de normalización para optar a una precisión mayor, no obstante, todo ello debe venir complementado con información sobre el análisis de la valoración, es decir, número de inmuebles utilizados, tipos de testigos empleados (compraventa, tasación o transacción real), y análisis de la desviación típica de μ.

Las ventajas que se encuentran en estos algoritmos es la fidelidad conceptual de la práctica de valoración por lo que resulta intuitivo y justificable, la selección adecuada de comparables y el bajo impacto de muestras deficientes. Sus inconvenientes principalmente vienen dados por la necesidad de contar con una gran cantidad de datos depurados para asegurar la calidad de estimación y su sensibilidad ante aquellas zonas más heterogéneas donde no se permite una homogeneización correcta del valor unitario.

2.1.1.3 Método KNN

El algoritmo KNN es similar en concepto al algoritmo de comparables puesto que ambos buscan elementos cercanos, los ponderan y les otorgan un valor. La diferencia es que la cercanía en este algoritmo viene dada por proximidad de todas las características calculada mediante la distancia euclídea, no únicamente el ámbito geográfico. Por lo tanto, resumiendo la explicación de Berásategui Arbeloa, Gonzalo (22 de marzo 2018) el proceso consta de obtener la muestra y seleccionar las variables de entrada, ponderar las variables para el cálculo de la distancia, determinar un número K de vecinos, calcular la relevancia de los vecinos según su distancia y obtener el valor. Las bondades de este algoritmo son su fácil justificación, su adaptabilidad y el control de fallos en muestras deficientes afectando

tan solo a sus vecinos. Sus inconvenientes, elevadas muestras necesarias y su coste computacional.

2.1.1.4 Redes neuronales artificiales

Morera Munt, Alba (febrero 2018) expone que el modelo del perceptrón multicapa es el más utilizado en el cálculo de regresiones. Se trata de un modelo bioinspirado cuyo funcionamiento se basa en la recepción por parte de las neuronas de unas señales de entrada del exterior a las cuales se le asignan unos pesos que se procesan mediante operaciones simples (no lineales) y resulta una señal de salida. Al tratarse de una red neuronal multicapa, existen múltiples capas de neuronas enlazadas entre sí con pesos asignados. Existen dos tipos de conexión, la total que implica que todas neuronas de salida de una capa son entrada de la siguiente y la parcial donde efectivamente como su nombre indica, la entrada a la siguiente capa es un subconjunto de neuronas. Otra capacidad de las redes neuronales es el ajuste progresivo por aprendizaje de los pesos de las conexiones con el fin de que se aproximen a la función objetivo. Finalmente, cada capa necesita una función de activación y entre las más empleadas se encuentran la función umbral, la gaussiana y la sigmoidal que se encargan de normalizar las variables de entrada.

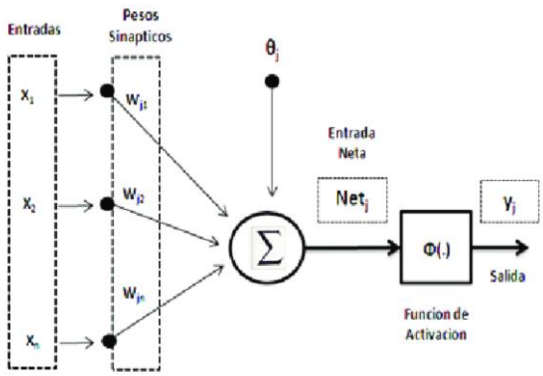

Figura 2-1 Estructura Red Neuronal[3]

Las ventajas vienen dadas por su capacidad de aprendizaje y su eficiencia computacional. Su desventaja principal es su difícil interpretación al ser un algoritmo de alta complejidad.

[3] https://www.researchgate.net/

2.1.1.5 Método hedónico

Estos métodos según Quiroga (2005) asumen que el valor de un inmueble es la suma del valor de todas sus características y que su agrupamiento por zonas geográficas aporta zonas homogéneas de valor. Los parámetros se calculan utilizando como variables de entrada características de dicho inmueble mediante regresión lineal o múltiple generalmente. Su principal ventaja es la sencillez del método pudiendo ser empleado por personal menos técnico y su inconveniente es la falta de detalle al tratarse de información agregada.

2.1.1.6 Random Forest

Random Forest fue desarrollado por Leo Breiman, los pormenores se detallan en la siguiente literatura, Breiman (1996).

Diertterich Thomas G. (2000) expone una serie de conceptos, en los cuales se encuentra que Random Forest es un ensemble de un conjunto de modelos base, lo que comúnmente se denomina como ensembles de tipo 'bagging'. El objetivo de este tipo de ensemble es construir varios modelos base en paralelo con el mismo algoritmo y clasificar los datos por votación. La idea es que, aunque un modelo base puede tener un mayor margen de error, es improbable que todos los modelos base tengan ese nivel de margen de error, si los errores están descorrelacionados. Es por tanto necesario que los modelos base sean diversos a fin de que los modelos base no erren de manera sistemática. La mejor práctica en este sentido es introducir cierta aleatoriedad en la construcción de modelos base.

Según Castillo (2015) la aleatorización es un proceso por el cual en vez de elegir el mejor atributo para un nodo se opta por escoger el mejor de un subconjunto de atributos en vez del total, con esto se consigue obtener un espectro más amplio y variado de modelos base que contemple todas las combinaciones posibles para que el aprendizaje sea más robusto. El método de Random Forest combina la idea de bagging de Breiman junto a la selección aleatoria de atributos, Ho (1995) y Amit et al (1997), para así conseguir una variación controlada.

Diertterich Thomas G. (2000) entre otros autores afirman que los ensembles cuentan con un gran desempeño incluso aunque los modelos base no sean muy precisos, siempre que dichos modelos base predigan mejor que el azar.

En este caso el modelo base se trata ni más ni menos del famoso y conocido árbol de decisión. Citando a Nuria Castillo, "Un árbol de decisión es una secuencia ordenada de preguntas en las que la siguiente pregunta depende de la respuesta a la pregunta actual. Dichas cuestiones son formuladas sobre las variables que definen cada elemento con el fin de acabar asignándoles una determinada clase. Este procedimiento, con sus correspondientes preguntas y bifurcaciones, es representado de forma natural mediante un árbol". Cada nodo del árbol representa un atributo, y cada rama un posible valor de este mismo. El primer nodo (raíz) es aquel que conecta con el resto de los nodos internos hasta llegar a un nodo hoja donde ya se ha conseguido clasificar la clase. Para la selección de atributos para cada nodo se optimizan los parámetros de la función de ganancia de la información expresada como:

$$I_j = H(S_j) - \sum_{i \in 1,2} \frac{|S_j^i|}{|S_j|} H(S_j^i)$$

donde S representa el conjunto de muestras que hay en el nodo a dividir, y S^i son los dos conjuntos que se crean de la partición. La función mide la entropía del conjunto, en caso de ser un problema de clasificación y funciones de distribución de probabilidad continuas en caso de ser uno de regresión como se muestra en la siguiente formula:

$$I_j = \sum_{v \in S_j} log(|\Lambda_y(v)|) - \sum_{i \in 1,2} \sum_{v \in S_j^i} log(|\Lambda_y(v)|)$$

donde (Λ_y) es la matriz de covarianzas.

Como se ha demostrado también existe la variante de regresión que es la que se va a utilizar en este estudio. Por lo que, en el caso de la regresión, el Random Forest completo se obtendría de la media aritmética de cada resultado final de cada árbol.

La selección de dicho algoritmo para este estudio se debe a un conjunto de razones, las cuales se explicarán a continuación, que hacen de este algoritmo una opción muy valiosa y robusta.

2.2 Critica del estado del arte

Las distintas metodologías que existen para la valoración utilizan características intrínsecas al inmueble. Esto se debe a la naturalidad e inmediatez de obtener estos datos, es decir, en caso de valorar un inmueble de ciertas características, se utilizan dichas características para la valoración.

En el caso del método de comparables, al comparar estas características con la de otros inmuebles, homogeneizarlas y realizar los cálculos para poder obtener un valor, a simple vista y también en la práctica, se observa que los modelos, tanto el de comparación, como el resto, son muy fiables y dan buenos resultados.

Por otra parte, cuando analizamos el cálculo realizado, se deduce un aislamiento del inmueble para con su entorno, lo cual se ha ido subsanando, incluyendo variables que describen el ámbito y la situación en la que se encuentra el inmueble. Un ejemplo de estas variables es la facilidad con la que se aparca en las inmediaciones, la cantidad de parques cerca, lo lejos que se encuentra las paradas de autobús o las estaciones de metro.

Aunque esta es la metodología establecida, se aprecia con la inclusión de las variables del entorno una mejora de varios puntos porcentuales en la precisión del modelo, y se suscita la duda de si el modelo ha alcanzado su máximo desempeño, o existe algún margen de mejora mediante la incorporación de nuevas variables, que describen algo más que las meras características del inmueble y de su entorno.

2.3 Propuesta

De una observación inicial y rápida se deduce el primigenio aislamiento del inmueble en cuanto a variables utilizadas para el modelo en los métodos de valoración previamente descritos, así mismo se observa una mejora con la inclusión de variables del entorno.

Como propuesta de mejora, se contempla la posibilidad de añadir al modelo variables que ayuden a describir el entorno económico en el que se encuentra, clasificada por su división geográfica. En concreto, se propone incluir una variable macroeconómica que recoge información agregada: el PIB. Esta variable es el candidato idóneo porque es una variable que se utiliza para determinar la producción total de un país en un momento determinado, en este caso específico se utilizará el PIB de cada comunidad autonómica, aportando así la división geográfica.

3 Diseño

3.1 Análisis del problema

3.1.1 Análisis de requisitos

Para determinar si la hipótesis expuesta anteriormente es verdadera o no, se debe ser capaz de reproducir un modelo de valoración automática de los usados en la actualidad, utilizando las mismas variables que definen los inmuebles y que se utilizan como parámetros actualmente (geolocalización, superficie, número de estancias, altura, etc). Posteriormente se debe incluir la variable macroeconómica elegida dentro de los parámetros para el cálculo del modelo, en este caso para ver el desempeño de la hipótesis y realizar una comparación entre ambos modelos, y con ello determinar si existe una mejora apreciable.

Los datos necesarios para la realización del modelo deberían ser de una fuente fiable que facilite los valores reales de una tasación, a su vez deberían incluir también las características de los inmuebles tasados. La calidad de los datos es sumamente importante, al igual que lo es la cantidad, por ello los datos deberían ser un número lo suficientemente representativo para poder realizar un modelo lo más robusto posible. Cabe resaltar que las tasaciones deberán haber sido acontecidas en un lapso no muy extendido en el tiempo (se aconseja no superior a un año).

La elección del modelo es una tarea posterior al análisis exploratorio, barajando tantas posibilidades como modelos existen en el mercado.

3.1.2 Análisis de soluciones

Este estudio se va a realizar con un conjunto de datos, obtenidos de una empresa privada española, dedica a las tasaciones de inmuebles de distinta tipología, con más de 25 años de experiencia. Este conjunto de datos inicial consta de exactamente 156367 tasaciones acontecidas entre el 1 de enero y el 31 de diciembre de 2019, ambos inclusive. Los datos están distribuidos en lo extenso del territorio nacional, lo que suscita la duda de si el modelo de futura creación será capaz, independientemente de cuál se elija, de contemplar todas las casuísticas que existen, dado que el set de datos engloba todas las transacciones de bienes inmuebles sin diferencias tipológicas. Hay que mencionar también que las condiciones socioeconómicas no son equivalentes en todos los rincones del país, es por

ello que se debe realizar una primera división provincial (listado de provincias en la tabla 3-1) para, por una parte, conocer el número de muestras por provincias y, posteriormente decidir cuáles son más o menos relevantes. Tras un primer análisis de los datos, en la Figura 3-1 se observa una clara mayoría de muestras en el código 28 perteneciente a la provincia de Madrid, seguida del código 08, perteneciente a Barcelona, pero con una diferencia de casi siete mil muestras. Además de ello, existen provincias con un número de muestras muy poco significativo, las cuales se descartan por no ser capaces de ofrecer una posible buena solución.

Figura 3-1. Distribución de muestras por provincias.

Para este primer estudio, se decide utilizar únicamente los datos correspondientes a la provincia de Madrid, dado que es la que más muestras tiene, y con la que posiblemente se obtengan mejores resultados en el momento de entrenar los modelos. Por otra parte, los datos macroeconómicos, que son el objetivo final de este estudio, son más estables en la línea temporal de esta área.

En cuanto a posibles modelos a desarrollar para la reproducción fiable de este problema, cabe mencionar que la mejor opción siempre será realizar una reproducción del modelo de comparables, dada su cercanía con la realidad. En este estudio no se dispone de datos suficientes para realizar un modelo de comparables fiable, es por ello que se procede a utilizar Random Forest, dado que es el segundo modelo que circunscribe las mayores posibilidades de obtener mejores resultados gracias a la robustez de su aprendizaje y que es fácilmente parametrizable, entre otras virtudes.

Código	Nombre	Código	Nombre	Código	Nombre
2	Albacete	16	Cuenca	36	Pontevedra
3	Alicante/Alacant	20	Gipuzkoa	26	Rioja, La
4	Almería	17	Girona	37	Salamanca
1	Araba/Álava	18	Granada	38	Santa Cruz de Tenerife
33	Asturias	19	Guadalajara	40	Segovia
5	Ávila	21	Huelva	41	Sevilla
6	Badajoz	22	Huesca	42	Soria
7	Balears, Illes	23	Jaén	43	Tarragona
8	Barcelona	24	León	44	Teruel
48	Bizkaia	25	Lleida	45	Toledo
9	Burgos	27	Lugo	46	Valencia/València
10	Cáceres	28	Madrid	47	Valladolid
11	Cádiz	29	Málaga	49	Zamora
39	Cantabria	30	Murcia	50	Zaragoza
12	Castellón/Castelló	31	Navarra	51	Ceuta
13	Ciudad Real	32	Ourense	52	Melilla
14	Córdoba	34	Palencia		
15	Coruña, A	35	Palmas, Las		

Tabla 3-1. Listado Provincias

3.2 Diseño de la solución

3.2.1 Análisis exploratorio

En este apartado se procederá a realizar un análisis de las variables que intervienen en la realización del modelo, para posteriormente identificar de entre todas las muestras, aquellas que son outliers.

El fichero inicial cuenta con las siguientes variables incluidas en la tabla 3-2, a continuación, se describen en mayor profundidad:

- Identificador, número automático auto incremental que identifica unívocamente a cada inmueble dentro del set de datos.
- Longitud y latitud, coordenadas geográficas utilizadas como sistema de referencia que permite la ubicación, únicamente horizontal, dado que la altura geográfica no se contempla en esta solución.
- Código postal, numeración de 5 dígitos, que representan las distintas zonas de un país y así facilitar la entrega de correo postales, los 2 primeros dígitos hacen referencia a las provincias, las otras 3 a zonas postales.
- Baños, número de baños/aseos o habitación destinada a ese propósito dentro de la vivienda.
- Habitaciones, número de habitaciones dentro el inmueble, son contabilizar el espacio destinado al salón.

- Superficie construida, área del polígono exterior delimitado por un espacio cubierto (suma de la superficie útil de la vivienda, y del espacio necesario para ubicar el cerramiento)
- Superficie parcela, es caso de tratarse de una vivienda unifamiliar, es el área del suelo libre de construcción.
- Antigüedad, cantidad de tiempo en años desde la construcción finalizada del inmueble hasta la fecha actual.
- Subtipología, subdivisión en distintos tipos dentro de las viviendas, siendo para este caso de uso simplificado una única división la que marque la diferencia (unifamiliar/plurifamiliar).
- altura, número de planta en la que está dispuesta la vivienda dentro del edificio.
- fecha tasación, fecha a la que se realizó la tasación.
- PIB, producto interior bruto de la comunidad a la que pertenece el inmueble según el trimestre en el que se realizó la tasación.
- Valor tasación, valor final del inmueble, presente en el informe de tasación presencial.

Atributo	Descripción
Identificador	identificador de la vivienda
Longitud	longitud de la vivienda
Latitud	latitud de la vivienda
CódigoPostal	código postal de la vivienda
Baños	número de baños/aseos de la vivienda
Habitaciones	número de habitaciones de la vivenda
SuperficieConstruida	superficie construida de la vivienda (en metros cuadrados)
SuperficieParcela	superficie de la parcela de la vivienda(en metros cuadrados) solo unifamiliares
Antigüedad	numero de años desde la construcción de la vivienda.
SubTipología	division de las viviendas en unifamiliar y plurifamiliar
Altura	altura de la vivienda dentro del edificio(para las viviendas plurifamiliares)
FechaTasacion	fecha de la tasacion de un inmueble
PIB	PIB por comunidades
ValorTasación	valor de la tasaacion de la vivienda

Tabla 3-2. Atributos de las muestras

El valor del campo PIB, se ha calculado según los datos publicados por la Comunidad de Madrid[4], se ha agrupado por trimestres con la ayuda del campo fecha de tasación y se ha

[4] http://www.madrid.org/iestadis/fijas/coyuntu/economicos/intermediocrm13.htm

autocompletado con los datos correspondientes. Antes de empezar con el análisis se procede a eliminar aquellos campos que no van a ser analizados dado que no van a ser usados dentro del modelo de forma explícita, tal como por ejemplo la fecha de tasación.

El set de datos inicial se divide en 2 grandes grupos característicos, unifamiliares y plurifamiliares, debido a que entre ellos tienen características más homogéneas y por tanto las vamos a analizar por separado. Para el caso de las viviendas plurifamiliares, se prescinde de la variable superficie de parcela, dado que no es relevante para esta tipología, por el contrario, se adopta la variable altura dentro del edificio.

Finalmente, las variables adoptadas para el estudio son: baños, habitaciones, superficie, antigüedad y altura. Veamos a continuación un análisis de estas variables en detalle, cuya descripción analítica se resumen en la tabla 3-3.

	Baños	Habitaciones	Superficie construida	Antigüedad	Altura
count	18924	18924	18924	18924	18924
mean	1,3	2,4	92,71	35,7	2,54
std	0,77	1,2	36,15	25,8	2,31
min	0	0	16	0	-3
25%	1	2	70	14	1
50%	1	3	87	38	2
75%	2	3	108	51	4
max	8	9	708	244	60

Tabla 3-3. Descripción analítica de atributos

3.2.1.1 Para viviendas plurifamiliares

Habitaciones, durante el análisis se observa en la figura 3-2 que existen viviendas de hasta 9 habitaciones, pero el grueso común se encuentra entre 0 y 4 habitaciones. En cuanto a la relación entre el precio y el número de habitaciones, se observa en la figura 3-3 que en su mayoría las viviendas no superan por lo general la barrera de los 300K euros, salvo alguna excepción. Se procede a considerar como outliers a todas las muestras que tengan más de 5 habitaciones.

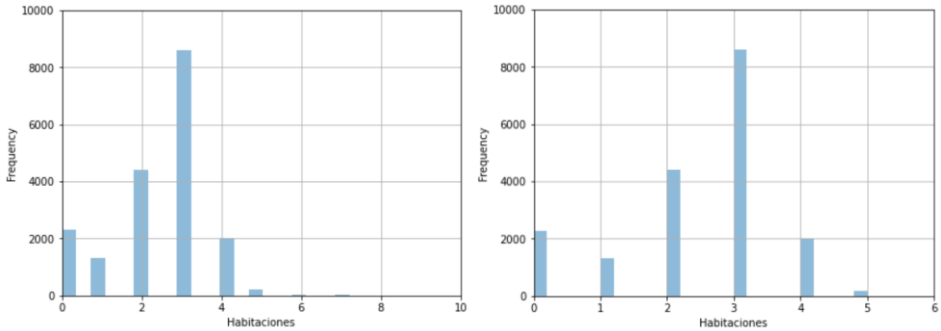

Figura 3-2. Comparativa de la distribución del número de habitaciones con y sin outliers.

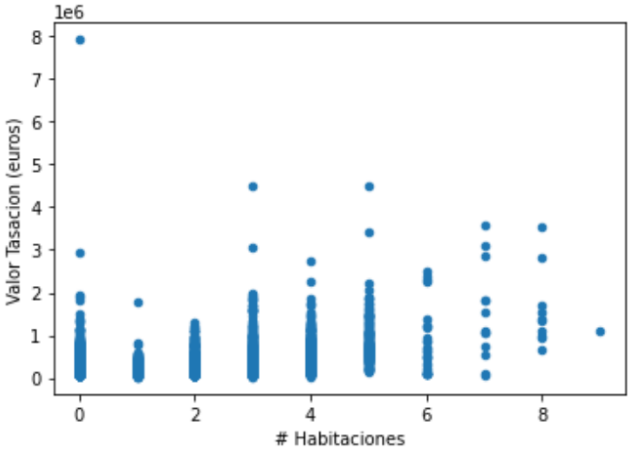

Figura 3-3. Dispersión del valor de tasación con respecto al número de habitaciones

Baños, en esta variable se observa en la figura 3-4 al igual que pasa con las habitaciones, que por lo general lo común se encuentra en una cifra inferior a 4, salvo alguna excepción, que sitúa a viviendas con hasta 8 baños. En cuanto a la relación entre el precio y el número de baños, se observa en la figura 3-5 que en su mayoría las viviendas no superan la barrera de los 300K euros, salvo un ligero repunte en viviendas con 4 baños, se consideran outliers todas aquellas viviendas que superen 4 cuartos de baño, por no ser una cantidad representativa.

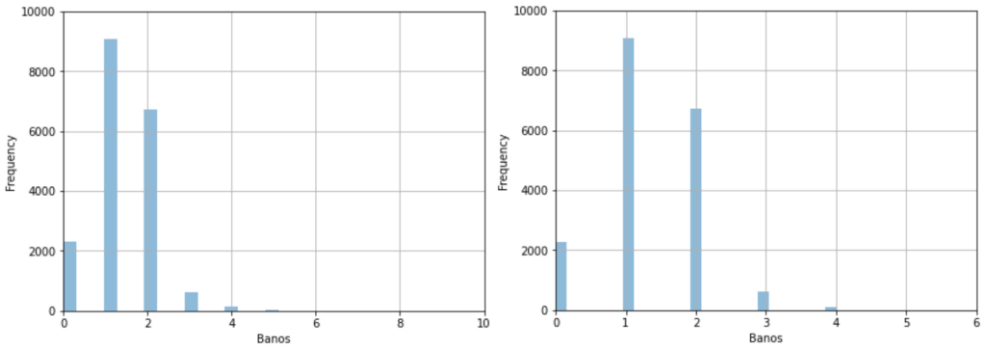

Figura 3-4. Comparativa de la distribución del número de baños con y sin outliers.

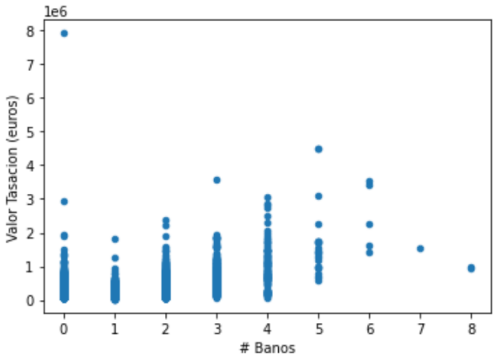

Figura 3-5. Dispersión del valor de tasación con respecto al número de baños.

Antigüedad, se observa que este parámetro tiene valores de hasta casi 250 años como muestra la figura 3-6. En el mercado actual de valoraciones automáticas para un modelo muy robusto se consideran outliers, aquellos inmuebles que sobrepasan los 200 años. En este caso de uso y viendo la distribución que tiene, se contabilizarán todos aquellos inmuebles únicamente con antigüedad inferior a 150, así mismo la figura 3-7 muestra que la dispersión con esos valores es estable con respecto al precio del inmueble.

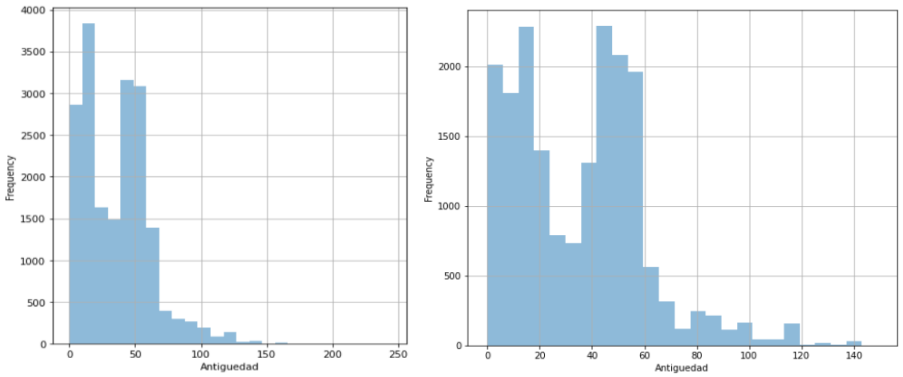

Figura 3-6. Comparativa de la distribución de antigüedad con y sin outliers.

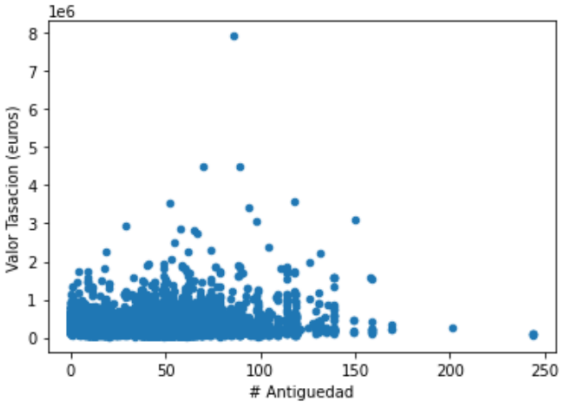

Figura 3-7. Dispersión del valor de tasación con respecto a la antigüedad del inmueble.

Altura, se observa que el grueso de las muestras se encuentra en edificios de hasta 12 plantas (figura 3-8), así mismo la relación que tiene con el valor del inmueble se mantiene estable y normal (figura3-9), en viviendas de poco más de 20 plantas. Es por ello por lo que, consideramos outliers todas aquellas viviendas que superen la planta 21.

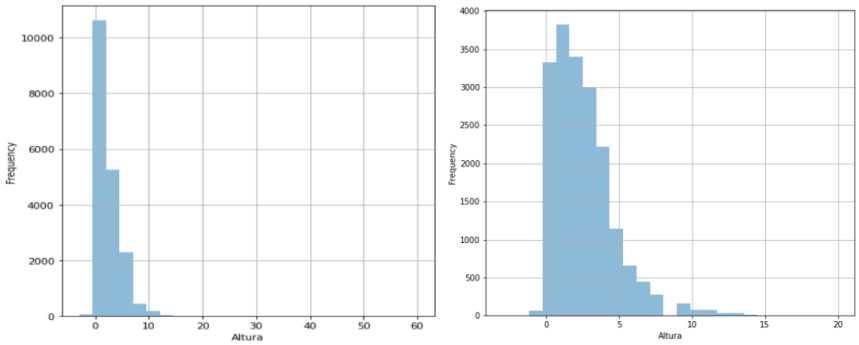

Figura 3-8. Comparativa de la distribución de la altura del inmueble con y sin outliers.

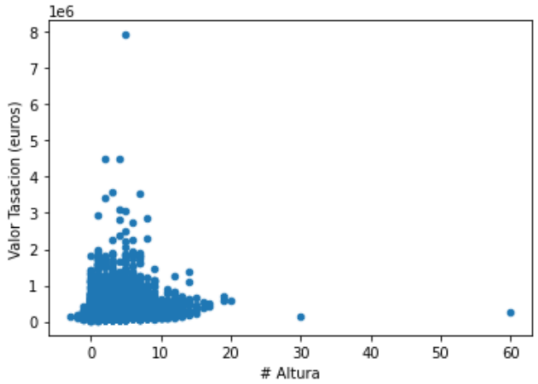

Figura 3-9. Dispersión del valor de tasación con respecto a la altura del inmueble.

Superficie construida, es sin duda la variable más representativa de cara al valor del inmueble. La casuística es diversa como observamos en la figura 3-10 con viviendas de hasta casi 700 metros cuadrados, en estos casos lo mejor es filtrar en una superficie máxima basada en conocimiento de negocio, que, por lo general para viviendas de tipo plurifamiliar, suele estar en el rango de 150 - 250 metros cuadrados (figura 3-11). En este análisis se procede a considerar como outlier toda aquella vivienda que supera los 250 metros cuadrados. En cuanto a la gráfica que representa la relación entre el precio de la vivienda y la superficie (figura 3-12) vemos que existe una clara correlación entre la superficie y el valor de la vivienda.

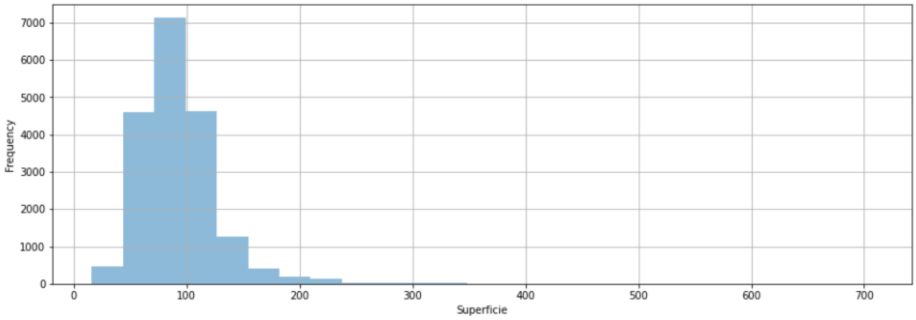

Figura 3-10. Distribución de la superficie del inmueble con outliers.

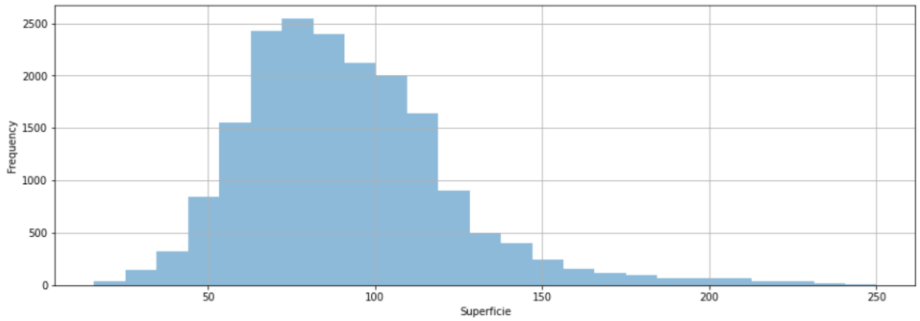

Figura 3-11. Distribución de la superficie del inmueble sin outliers.

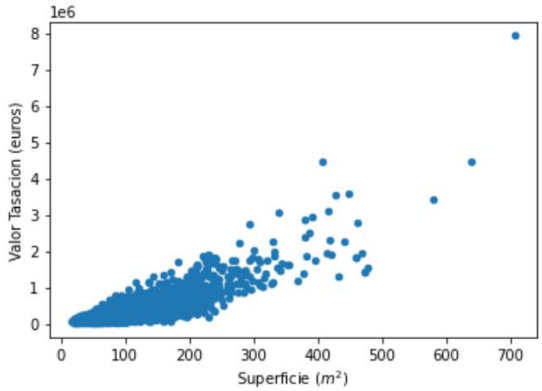

Figura 3-12. Dispersión del valor de tasación con respecto a la altura del inmueble.

3.2.1.2 Para viviendas Unifamiliares

Habitaciones, durante el análisis se observa en la figura 3-13, que existen viviendas de hasta 9 habitaciones, pero el grueso común se encuentra entre 0 y 6 habitaciones. En caso de viviendas de tipología unifamiliares se observa que el número de habitaciones puede encontrarse dentro de 0 y 9 de forma común, es por ello por lo que no consideramos ningún caso como outlier. La dispersión del valor tiene una incidencia en viviendas de 1 habitación como muestra la figura 3-14.

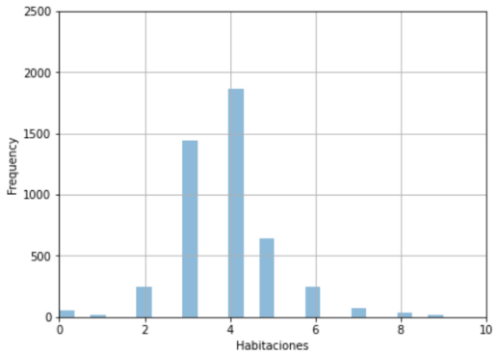

Figura 3-13. Distribución del número de habitaciones del inmueble

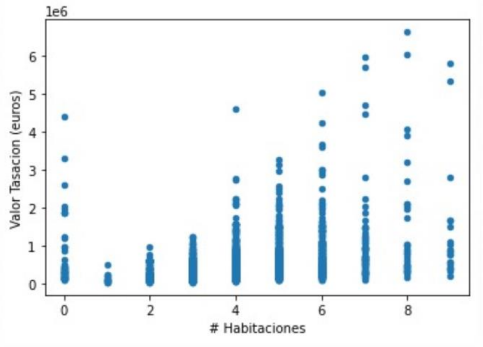

Figura 3-14. Dispersión del valor de tasación con respecto al número de habitaciones.

Baños, en esta variable se observa al igual que pasa con las habitaciones en las viviendas unifamiliares, que por lo general lo común se encuentra en una cifra inferior a 7 (figura 3-15), las viviendas con un número superior no pueden considerarse outliers, porque la relación que existe entre el valor de la vivienda y el número de baños sigue la misma tendencia que en proporciones menores (figura 3-16) y el número de cosos existentes es significativo.

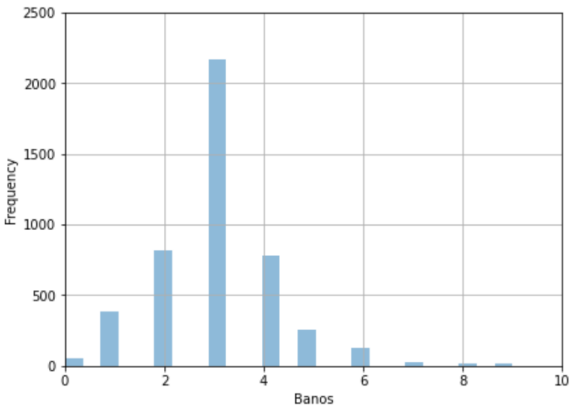

Figura 3-15. Distribución del número de baños del inmueble

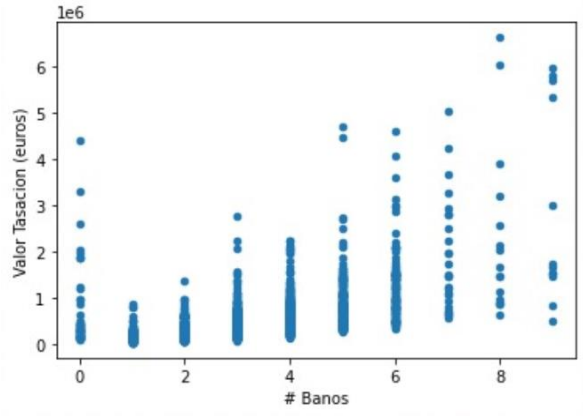

Figura 3-16. Dispersión del valor de tasación con respecto al número de baños.

Antigüedad, se observa que este parámetro tiene valores de hasta casi 200 años(figura 3-17). En el mercado actual de valoraciones automáticas para un modelo muy robusto se consideran outliers, aquellos inmuebles que sobrepasan los 200 años, basándose en conocimientos adquiridos a lo largo de la experiencia laboral en este caso de uso y viendo la distribución que tiene (figura 3-18), se contabilizarán todos aquellos inmuebles únicamente con antigüedad inferior a 150, como ocurría en el caso de las viviendas plurifamiliares.

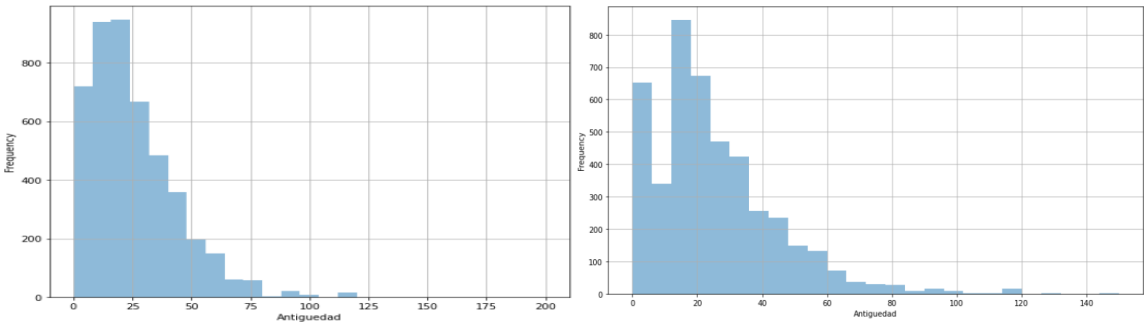

Figura 3-17. Comparativa de la distribución de la antigüedad del inmueble con y sin outliers.

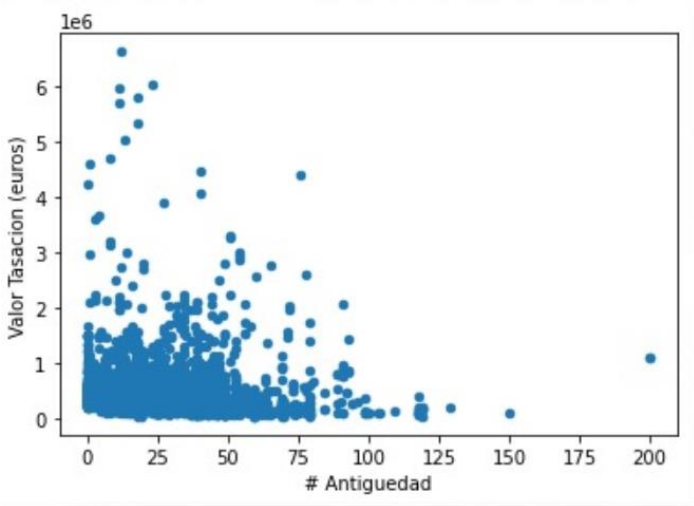

Figura 3-18. Dispersión del valor de tasación con respecto a la antigüedad del inmueble.

Superficie parcela, esta variable es utilizada únicamente en la tipología de viviendas unifamiliares, debido a su gran implicación de cara al valor final del inmueble. Se observa que el grueso de las muestras se encuentra en superficies de parcela inferiores a 5000 metros cuadrados (figura 3-19). Al igual que en el caso de la antigüedad se procede a utilizar un valor comúnmente usado en el mercado actual, que se ajusta de forma acertada a la realidad. Por ello todas las viviendas con superficie de parcela superior a 2500 metros cuadrados, quedarán excluidos de set final de datos (figura 3-20).

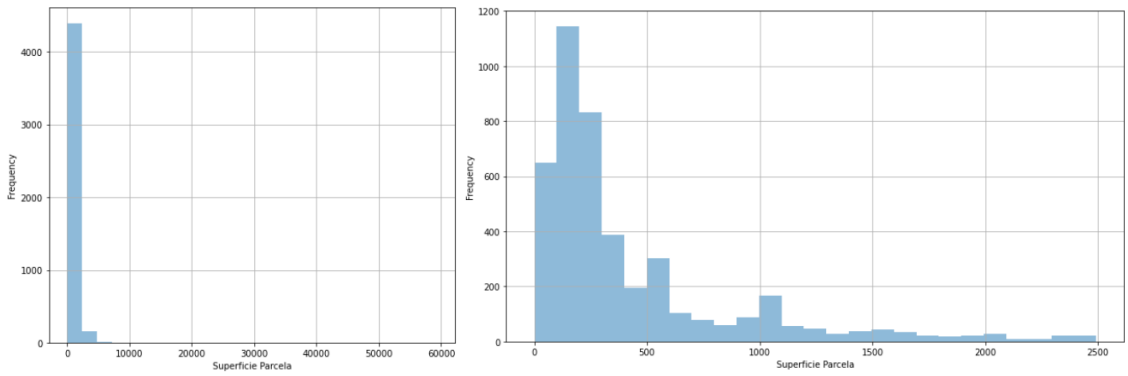

Figura 3-19. Comparativa de la distribución de la superficie de parcela con y sin outliers.

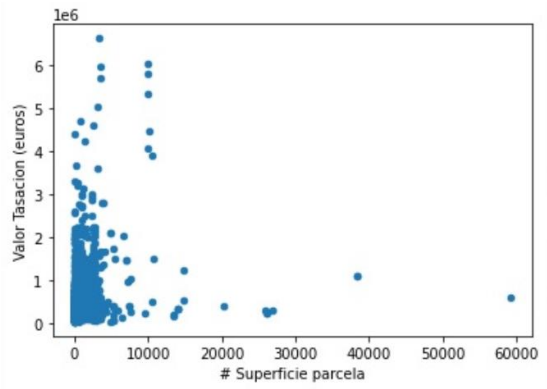

Figura 3-20.Dispersión del valor de tasación con respecto a la antigüedad del inmueble.

Superficie construida, es de nuevo la variable más representativa de cara al valor del inmueble. El conjunto de datos muestra unas superficies de hasta 20000 metros cuadrados (figura 3-21) la casuística es diversa y en estos casos lo mejor es filtrar en una superficie máxima basada en la experiencia y el conocimiento adquirido en estos últimos 2 años de experiencia trabajando en una tasadora, que por lo general para viviendas de tipo unifamiliar, suele estar en el rango de 150 - 650 metros cuadrados (figura 3-22), en este análisis se procede a considerar como outlier toda aquella vivienda que supera los 750 metros cuadrados, el gráfico de dispersión (figura 3-23), vemos la relación que existe entre valor final de la vivienda y la superficie, es casi lineal.

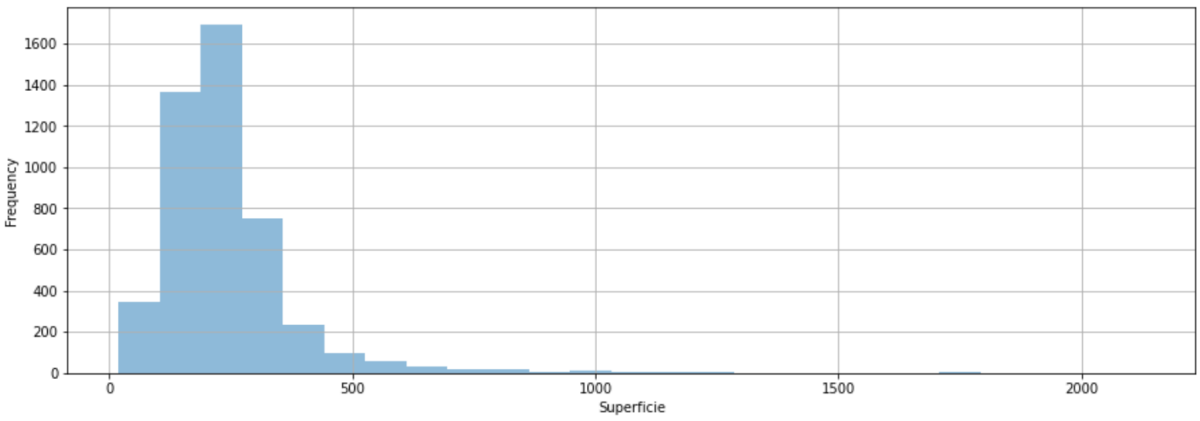

Figura 3-21. Distribución de la superficie del inmueble con outliers.

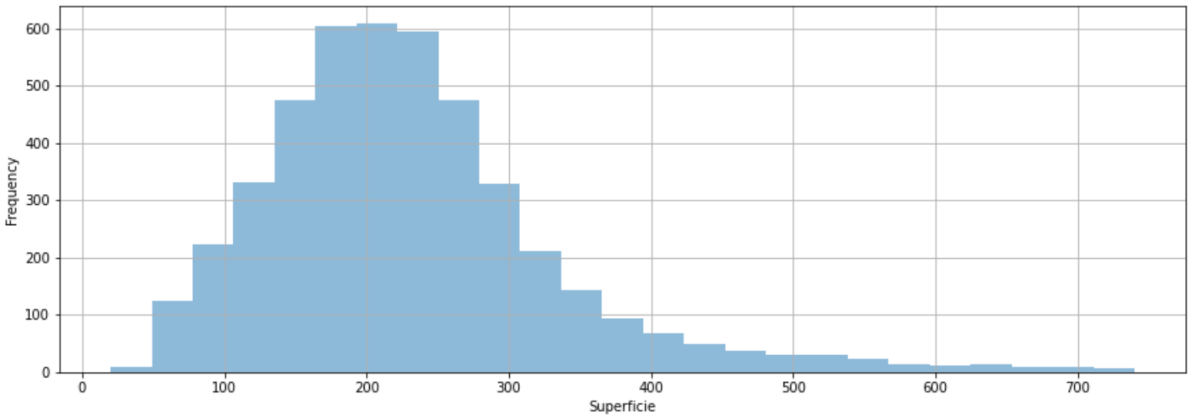

Figura 3-22. Distribución de la superficie del inmueble sin outliers.

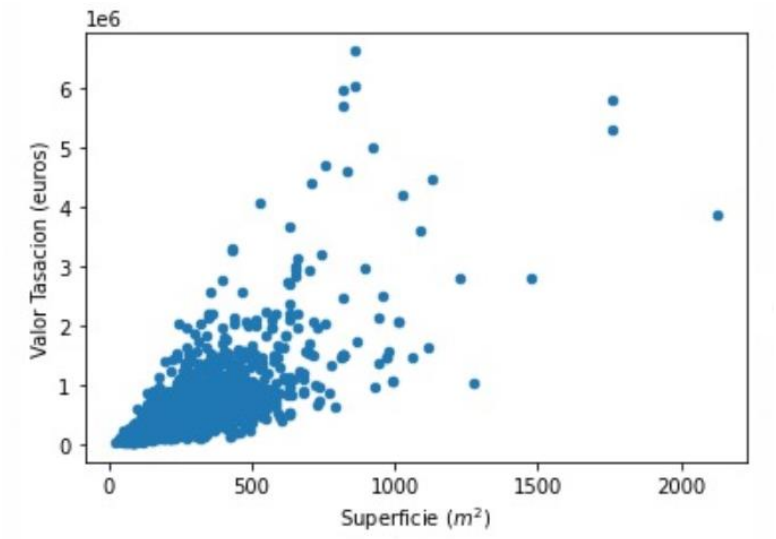

Figura 3-23. Dispersión del valor de tasación con respecto a la superficie del inmueble.

3.2.2 Diseño del Software

En este apartado se explica el análisis y las decisiones tomadas en el diseño del software.

Las opciones de lenguaje de programación barajadas para la implementación fueron R y Python, siendo este último el elegido finalmente, debido a que R dispone de un enfoque más matemático. Por otra parte, Python tiene un enfoque multipropósito y dado que el análisis visto en este informe no tiene tanta complejidad matemática, Python permite tener una homogeneidad en el resto del código (no necesariamente de análisis).

Los datos de las muestras utilizadas vienen en formato csv.

El diseño consta de un notebook, para el análisis exploratorio y un script que genera los modelos y evalúa los resultados y que se van a detallar en la siguiente sección.

4 Desarrollo

4.1 Implementación

La implementación final del proyecto se realiza con la ayuda de Python en su versión 3.6 para la creación de los modelos en la modalidad scripting. También se hace uso de notebooks de Jupyter para el análisis exploratorio y de resultados. Nótese de nuevo que los datos se obtienen directamente de ficheros csv.

En este apartado se explica la estructura del código implementado:

- **house_rf.py**, script diseñado para la generación de los modelos, hace uso de las siguientes librerías de Python:
    - *pandas*, biblioteca para la manipulación y análisis de datos, ofrece estructuras y operadores que facilitan las tareas relacionadas.
    - *numpy*, biblioteca que soporta vectores y matrices, al igual que las operaciones entre ellos.
    - *pickle*, biblioteca que permite el almacenamiento y carga de modelos.
    - *argparse*, biblioteca que permite procesar argumentos y opciones de línea de comando.
    - *sklearn*, biblioteca para el aprendizaje automático.

En el script se incluyen los siguientes métodos:

- get_random_params, genera un objeto con los parámetros de entrenamiento aleatorio, o se generan a partir de un fichero. El resultado se utiliza para realizar una búsqueda aleatoria de parámetros generales.
- get_grid_params, genera un objeto con los parámetros de entrenamiento, basado en el resultado del método anterior, o genera un set de parámetros partiendo de un fichero, utilizado para realizar una búsqueda exhaustiva de parámetros.
- calculate_ramdomized_search, realiza la búsqueda aleatoria de hiperparametros.
- calculate_grid_search, realiza la búsqueda exhaustiva de parámetros.
- config_data, prepara los datos para la realización del modelo.
- split_data, subdivide los datos para el entrenamiento y el test.
- create_model, crea el modelo con todos los parámetros previamente facilitados.

- evaluate, realiza los cálculos necesarios para medir el desempeño del modelo
- rs_evaluation, realiza los cálculos necesarios para medir el desempeño del modelo según la metodología utilizada en el sector del Real Estate.

Las opciones aceptadas por el script se pueden consultar en el siguiente gráfico.

```
usage: house_rf.py [-h] [-f < data_file.csv >] [-v < rs >] [-p] [-ss]
                   [-ds < 0.2 >] [-rs] [-f_rs < random_params.json >]
                   [-n_rs < 100 >] [-fo_rs < random_best_params.json >] [-gs]
                   [-f_gs < random_best_params.json >] [-cv_gs < 100 >]
                   [-fo_gs < grid_best_params.json >]
                   [-fo_gs_m < best_model.pkl >] [-nm]
                   [-f_nm < model_params.json >] [-s]
                   [-fo_nm < best_model.pkl >] [-em]
                   [-f_em < best_model.pkl >]

optional arguments:
  -h, --help            show this help message and exit
  -f < data_file.csv >  Data file name
  -v < rs >             House type: uni plu
  -p                    Procesar con macrovariables
  -ss                   Standart Scaler
  -ds < 0.2 >           Data Split: Test size [0-1]
  -rs                   Random_Search
  -f_rs < random_params.json >
                        json file name for random params
  -n_rs < 100 >         n iters for Random Search
  -fo_rs < random_best_params.json >
                        json file name for random params
  -gs                   Grid_Search
  -f_gs < random_best_params.json >
                        Data file name
  -cv_gs < 100 >        n cv folds for Grid Search
  -fo_gs < grid_best_params.json >
                        json file name for best grid params
  -fo_gs_m < best_model.pkl >
                        file name for best model
  -nm                   New Model
  -f_nm < model_params.json >
                        Data file name
  -s                    Save model
  -fo_nm < best_model.pkl >
                        New model file name
  -em                   Evaluate Model
  -f_em < best_model.pkl >
                        Model file name
```

Figura 4-1. Descripción de opciones del script

El Código del análisis descriptivo, es muy simple, en él se organizan los datos, en 2 grupos, por un lado, el conjunto de viviendas plurifamiliares, por otro lado el de unifamiliares, posteriormente se grafican los datos para analizarlos y realizar la limpieza, de datos incompletos, analizar los datos atípicos, y finalmente filtrar lo que será útil para el modelo.

5 Integración, pruebas y resultados

Se procede a realizar la integración del código implementado con los datos en formato csv y a continuación se describen las pruebas y los resultados de esas pruebas, finalmente se procede a capturar los resultados del entrenamiento en gráficos para que sean fácilmente interpretables.

El set inicial de datos cuenta con 23.166 datos inicialmente, los cuales se procede a dividir por tipología, por un lado, viviendas unifamiliares (4.399) y plurifamiliares (18.767). Para cada uno de estos subconjuntos, se realiza las siguientes operaciones para la creación del modelo:

1. **Preproceso**, División del conjunto de datos en 2 subconjuntos, uno para el entrenamiento (80%) y el otro para la verificación o test (20%), la división es aleatoria.

2. **Búsqueda de hiperparametros**, esta búsqueda se realiza para encontrar los parámetros de configuración del modelo, que arrojen un mejor desempeño. Estos parámetros, incluyen el número de árboles, el número de muestras en cada hoja divisoria, el número de muestras en cada hoja final, la profundidad y el número de atributos. Clarificar que este proceso se etiqueta como un prueba-error más que un problema de teoría dado que encontrar la configuración óptima requiere tiempo. También cabe mencionar que en la búsqueda de hiperparámetros se utiliza la técnica Cross Validation[5], que permite evitar problemas se sobreajuste[6], y en este caso la implementación utilizada ha sido 5-Fold. Así mismo se desconoce a priori la configuración de parámetros idónea para el modelo, es por ello que partiendo de una vaga idea se seleccionan un conjunto de parámetros, acorde al número de variables del problema, de tal forma que se obtiene un conjunto combinaciones, las cuales se proceden a evaluar en el desempeño del modelo, en este caso de uso según los parámetros seleccionados el número de combinaciones se elevan a 56000,

[5] https://es.wikipedia.org/wiki/Validaci%C3%B3n_cruzada

[6] https://es.wikipedia.org/wiki/Sobreajuste

que si se multiplican por el número de K particiones seleccionadas anteriormente, el número asciende a 280.000 evaluaciones de modelos, es por ello que el proceso se divide en 2 fases:

La primera, consiste en realizar es una búsqueda aleatoria de hiperparámetros, partiendo de las combinaciones anteriores de los parámetros previamente seleccionados, y realizando una validación cruzada para cada caso seleccionado. Este proceso se realiza un número suficiente de iteraciones (no tan grande como el numero inicial de posibles validaciones, pero sí lo suficientemente representativo), en este análisis se optó por realizar el equivalente al 10% del total (28000 iteraciones). Como resultado de este proceso se obtiene la configuración más optima, entre las evaluadas, este proceso ayuda a reducir coste de tiempo y recursos.

La segunda fase consiste en realizar una búsqueda completa de todas las combinaciones de parámetros facilitados, para ello, se toma como muestra inicial los parámetros obtenidos en la fase 1, y se concentra el esfuerzo alrededor de esos parámetros explícitamente, esta fase realiza al igual que la anterior una validación cruzada para poder localizar aquella combinación que dé el mejor desempeño del modelo. Como resultado de este proceso se obtienen los parámetros con lo que el modelo realizará su mejor desempeño.

3. **Creación del modelo**, como se ha observado, la fase de búsqueda de hiperparámetros, finaliza con la obtención de unos parámetros de configuración específicos, los cuales son sin duda, los que hacen del desempeño del modelo, los mejores de entre todas las combinaciones. Estos parámetros (figura 5-1) son los utilizados para la creación del modelo final y, como se ha indica anteriormente, el conjunto de datos destinado para el entrenamiento es el 80% de los datos del conjunto inicial, seleccionados de forma aleatoria, mientras el 20% restantes se utilizan para realizar las pruebas de validación.

   El número total de modelos generados es 4, unifamiliar con PIB incluido, unifamiliar sin PIB, plurifamiliar con PIB, plurifamiliar sin PIB; todos implementados con la misma metodología, y en cada caso los parámetros de configuración se han adaptado a cada problema.

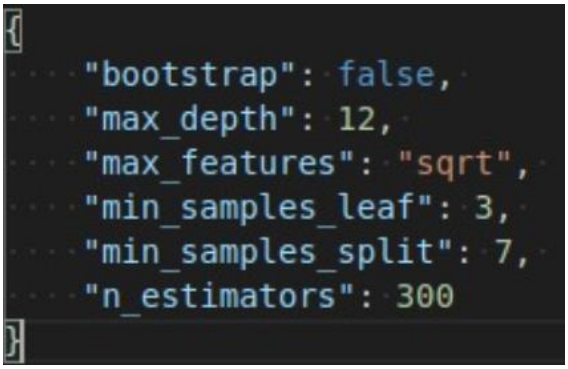

Figura 5-1. Ejemplo de parámetros configuración modelo

4. **Evaluación del modelo**, Se realiza la evaluación de los 4 modelos desarrollados, utilizando los siguientes métodos de evaluación:

   Error absoluto medio (MAE / EAM)[7], medida de la diferencia entre el valor real y el obtenido, cuya formula esta descrita en la figura 5-2

$$\mathrm{EAM} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} = \frac{\sum_{i=1}^{n} |e_i|}{n}.$$

Figura 5-2. Fórmula Error Absoluto Medio

   Error porcentual absoluto medio (MAPE)[8], medida de precisión de la predicción, cuya formula es la descrita en la figura 5-3

$$\mathrm{M} = \frac{100\%}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right|$$

Figura 5-3. Fórmula Error porcentual absoluto medio

[7] https://es.wikipedia.org/wiki/Error_absoluto_medio

[8] https://es.qwe.wiki/wiki/Mean_absolute_percentage_error

Precisión[9], medición de la dispersión del conjunto de valores obtenidos a menor dispersión mayor precisión, es equivalente a 1 – MAPE.

Precisión RE, al igual que la precisión del modelo, descrita anteriormente, en el sector inmobiliario, existe una medida de precisión que consta de etiquetar un valor como aceptable, si se encuentra dentro del 20% de error. Es decir que todos aquellos valores que estén en este rango se consideran buenos.

Para los modelos unifamiliares se obtienen los datos descritos en la tabla 5-1, donde se incluyen, ambos modelos con y sin PIB. Para compararlos, se toma como referencia la precisión RE que es la más utilizada en el sector inmobiliario, y se observa una mejora de casi 3 puntos de cara al conjunto de datos de entrenamiento para el modelo con variables macroeconómicas incluidas, así mismo poco más de 4 puntos porcentuales para el conjunto de datos de prueba.

	Unifamiliar sin Macroeconómicas		Unifamiliar con Macroeconómicas	
Medidas	Entrenamiento	Test	Entrenamiento	Test
Error Medio Absoluto (MAE)	47851,6	46929,38	41774,33	37437,41
Error porcentual absoluto medio (MAPE)	11,93%	12,14%	10,66%	10,06%
Precisión	88,07%	87,86%	89,34%	89,94%
Precisión Real Estate	84,01%	83,07%	86,91%	87,61%

Tabla 5-1. Resultados evaluación modelos Unifamiliares

Para plurifamiliares se obtiene los datos detallados en la tabla 5-2, donde se incluyen ambos modelos. Comparando ambos modelos, se toma como medida de referencia el mismo que anteriormente, la precisión RE, y se observa una mejora de algo más de 5 puntos porcentuales de cara al conjunto de datos de entrenamiento para el modelo con variables macroeconómicas incluidas, así mismo poco más de 2,5 puntos porcentuales para el conjunto de datos de prueba.

[9] https://es.wikipedia.org/wiki/Precisi%C3%B3n_y_exactitud

	Plurifamiliar sin Macroeconómicas		Plurifamiliar con Macroeconómicas	
Medidas	Entrenamiento	Test	Entrenamiento	Test
Error Medio Absoluto (MAE)	25843,14	26169,74	22850,56	23480,17
Error porcentual absoluto medio (MAPE)	12,94%	13,27%	11,23%	11,91%
Precisión	87,06%	86,73%	88,77%	88,09%
Precisión Real Estate	80,76%	80,05%	85,09%	82,69%

Tabla 5-2. Resultados evaluación modelo Plurifamiliares

5. **Evaluación Final,** se ha realizado una prueba final que consiste en evaluar los modelos obtenidos con datos más recientes (enero 2020)[10], el total de muestras del conjunto es 2.078, de los cuales 403 pertenecen a viviendas unifamiliares, los restantes 1.675 pertenecen a viviendas plurifamiliares. Consideramos esta prueba como un caso más cercano a la realidad. Dado que no se tiene ninguna muestra de ese año dentro del conjunto de datos utilizados para el desarrollo del modelo. Se observa en el detalle del modelo, que la mejora se mantiene para datos con fechas posteriores, se ha reducido la precisión, pero se sigue manteniendo por encima de 1.5 puntos porcentuales, tal y como muestra la tabla 5-3, para el caso de viviendas plurifamiliares, Por otro lado, en la figura 5-4 se incluye la distribución del error, para el caso del modelo que no contempla variables económicas, así mismo en la figura 5-5 se incluye el otro modelo que si cuenta con la variable macroeconómica, ambas igualmente para el caso de viviendas plurifamiliares. Para las viviendas unifamiliares, se presenta los detales en la tabla 5-4, así mismo la distribución de los errores para el modelo sin inclusión de variables macroeconómicas (figura 5-6), y al igual que el modelo que si lo hace en la figura 5-7. Con una mirada rápida a la precisión, se observa que se mantiene la mejora de valor para el modelo que incluye variables macroeconómicas en comparación a la que no lo hace, esta mejora es de lago más de 2 puntos porcentuales.

[10] El modelo actual se ha entrenado con datos de 2019.

Medidas	Plurifamiliar sin Macroeconómicas	Plurifamiliar con Macroeconómicas
Error Medio Absoluto (MAE)	46595,38	41704,76
Error porcentual absoluto medio (MAPE)	17,03%	15,15%
Precisión	82,97%	84,85%
Precisión Real Estate	82,27%	83,70%

Tabla 5-3. Detalle Evaluación Final Plurifamiliar

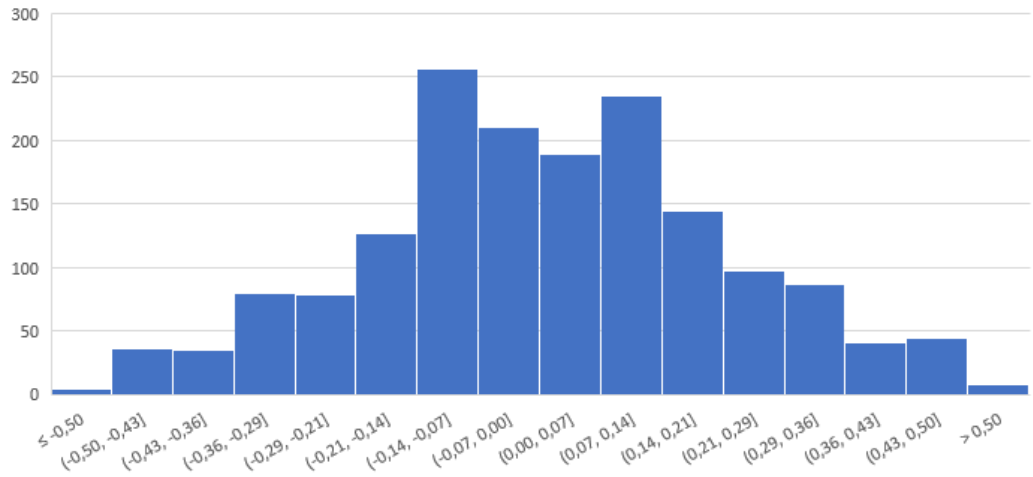

Figura 5-4. Distribución Error (Plurifamiliar) modelo sin V. Macroeconómicas.

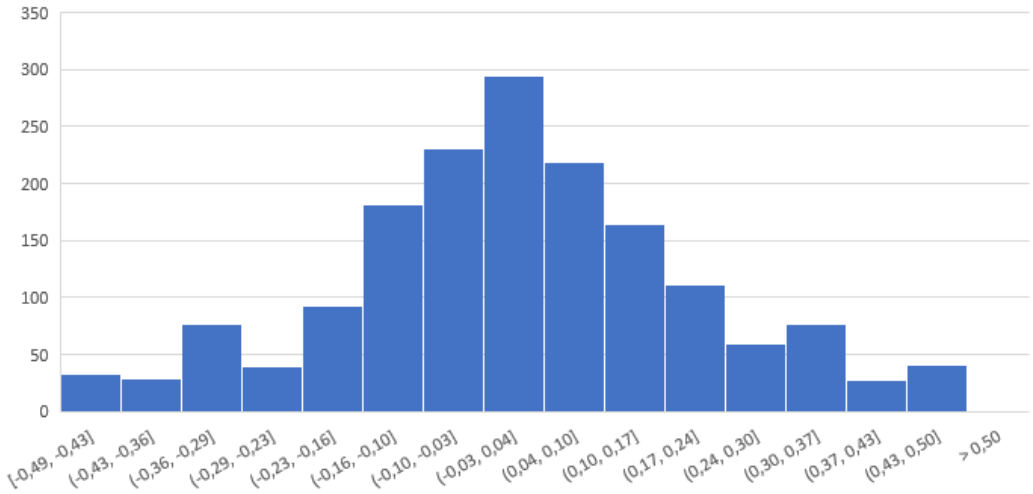

Figura 5-5. Distribución Error (Plurifamiliar) modelo con V. Macroeconómicas.

Medidas	Unifamiliar sin Macroeconómicas	Unifamiliar con Macroeconómicas
Error Medio Absoluto (MAE)	78336,49	70511,62
Error porcentual absoluto medio (MAPE)	16,94%	15,14%
Precisión	83,06%	84,86%
Precisión Real Estate	83,37%	85,61%

Tabla 5-4. Detalle Evaluación Final Unifamiliar

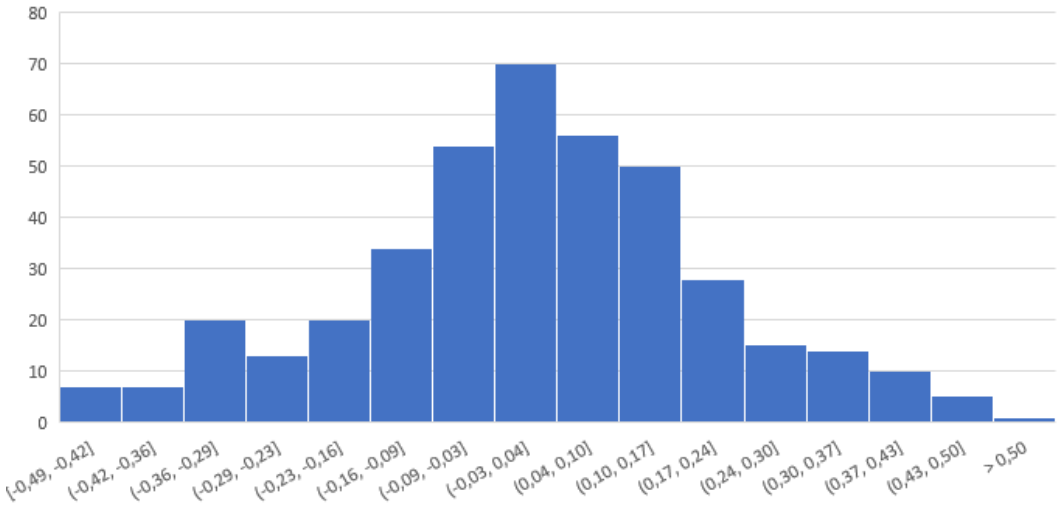

Figura 5-6. Distribución Error (Unifamiliar) modelo sin V. Macroeconómicas.

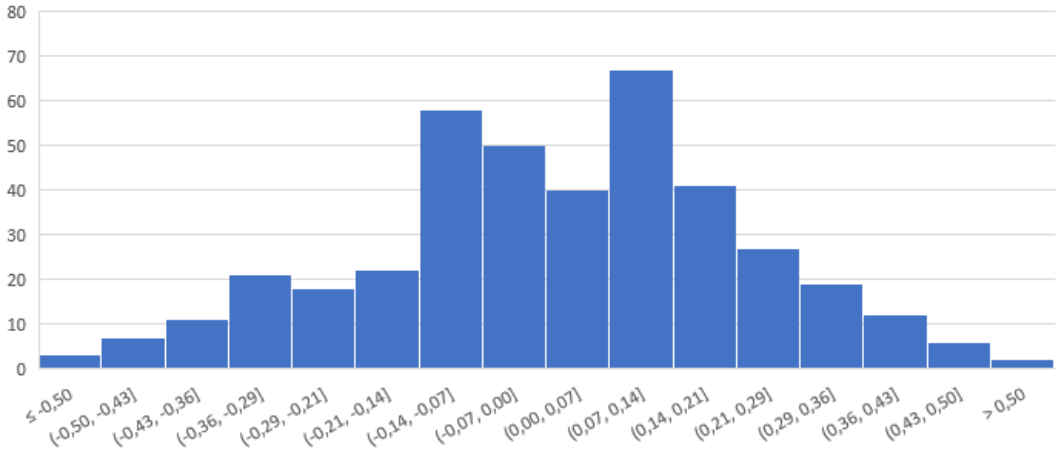

Figura 5-7. Distribución Error (Unifamiliar) modelo con V. Macroeconómicas.

6 Conclusiones y trabajo futuro

6.1 Conclusiones

Haciendo una recapitulación de la hipótesis inicial, el estudio concluido pretendía llegar a mostrar lo favorable que es incluir variables macroeconómicas en los modelos de valoración.

Se partió de una consigna muy sencilla, que el modelo primigenio de valoración automática solo se basaba en variables que eran características que describían el inmueble de forma aislada. La Asociación Española del Valor AEV (2019), en su documentación oficial, recomienda el uso de variables del entorno de la vivienda, las cuales se han ido integrando en los actuales modelos del mercado, tal como refiere el equipo de trabajo encargado de las valoraciones automáticas de la empresa idealista, AVM-Idealista (2019), otro ejemplo de estas mejoras las refiere el equipo de consultoría y valoración de gesvalt, Equipo de Investigación Gesvalt (2018) en sus documentos oficiales.

Se planteo la duda de si los modelos actuales del mercado estaban tocando techo, o si existía un margen de mejora, con la inclusión de más variables del entorno u otras variables. Como opciones se barajó la inclusión de variables económicas, decantando la balanza por una macroeconómica, ya que esta recoge información agregada y a su vez es una foto rápida de la economía actual de una región.

La implementación de un modelo para la provincia de Madrid utilizando un Random Forest ha sido positiva debido a su robustez y facilidad de parametrización.

Observando los resultados finales realizados en el conjunto de 2078 inmuebles ajenos al conjunto de datos inicial se puede observar una mejoría por parte del modelo que incluye variables macroeconómicas con respecto al que no, demostrándose por tanto la hipótesis inicial de este trabajo y abriendo camino a nuevas investigaciones en esta línea.

6.2 Trabajo futuro

Se ha observado con el análisis realizado, que los modelos mantienen una evolución constante, desacelerada en el tiempo, pero continuada.

Para los siguientes pasos, se propone realizar un análisis más amplio de detección de outliers, dado que se ha visto durante el estudio acometido, que las necesidades de viviendas van cambiando con el pasar del tiempo y esto es una razón más para estar atento a los cambios, decir con esto que es probable que basarse únicamente en un análisis estadístico, puede incluso perjudicar el estudio realizado, incluir la opinión de un experto o realizar una segregación por zonas, para determinar con mayor precisión los valores atípicos, también es una opción, tener en cuenta que no todas las regiones de un país son homogéneas.

Por otra parte, en ciertas ocasiones tendrá mucha relación la variable macroeconómica que se elija para la mejora del modelo con la región que se está estudiando, es por ello que la elección de dicha variable merece un análisis exhaustivo basándose en la región de estudio, la población y la economía propiamente.

De cara a los modelos utilizados, la comparativa entre varias implementaciones también tendrá connotaciones positivas, dado que la falta de muestras suficientes en ciertas regiones se pueda solventar con otras metodologías, de forma contraria, el exceso de datos podría ser perjudicial por un sobre entrenamiento.

Referencias

[1] AEV (2 de julio 2019). Estándar sobre valoración de inmuebles mediante modelos automatizados (AVM).

[2] Amit, Y., & Geman, D. (1997). Shape Quantization and Recognition with Randomized Trees. Neural Computation, 9(7), 1545–1588. https://doi.org/10.1162/neco.1997.9.7.1545

[3] AZNAR-BELLVER, J., et al (2012). Valoración inmobiliaria. Métodos y aplicaciones. España e Iberoamérica. Valencia: Editorial Universitat Politècnica de València

[4] Berásategui Arbeloa, Gonzalo (22 de marzo 2018). Implementación del algoritmo de los k vecinos más cercanos (k-NN) y estimación del mejor valor local de k para su cálculo. Grado en Ingeniería Informática. Supervisado por Mikel Galar Idoate.

[5] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32

[6] Castillo González. Nuria Victoria (septiembre 2015). Tecnicas de Machine Learning para el Post-Proceso de la predicción de la Irradiancia. Máster en Estadística Aplicada. Supervisado por D. Ramón Gutiérrez Sánchez.

[7] Consultoria y Valoración Gesvalt (2020) Valoración automática de viviendas.

[8] Dander Sánchez, O.A. (2012). Historia de la arquitectura I. RED TERCER MILENIO.

[9] Diertterich Thomas G. (2000). An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization. Kluwer Academic Publishers. Machine Learning, vol 40, pp. 139–157.

[10] Equipo AVM Idealista (2020) https://www.idealista.com/data/productos/valoracion-de-inmuebles/avm

[11] Ho, T. K. (1995). Random Decision Forest. Retrieved from http://cm.bell-labs.com/cm/cs/who/tkh/papers/odt.pdf

[12] Jerónimo AZNAR (2016) Valoración de Viviendas en España. El método de Homogeneización y metodologías alternativas. Finance, Markets and Valuation. Universidad Politecnica de Valencia.

[13] Julio Gallego Mora-Esperanza (2018) Modelos de valoración automatizada.

[14] Marketing y Tecnología (5 de septiembre, 2019). La economía colaborativa en el sector inmobiliario lidera las nuevas tendencias Proptech. Inmodiario. https://www.inmodiario.com/197/28007/economia-colaborativa-sector-inmobiliario-lidera-nuevas-tendencias-proptech.html

[15] Morera Munt, Alba (febrero 2018). Introducción a los modelos de redes neuronales artificiales. El Perceptrón simple y multicapa. Grado en Matemáticas. Supervisado por Tomás Alcalá Nalvaiz

[16] Quiroga, Bernardo (2005). Precios hedónicos para valoración de atributos de viviendas sociales en la región metropolitana de Santiago. Pontificia universidad Católica de Chile. Facultad de ciencias económicas y administrativas instituto de economía

[17] Tinsa (15 febrero, 2016). Tasación de una vivienda, los motivos de su importancia. https://www.tinsa.es/blog/tasaciones/el-valor-de-tasacion-de-una-vivienda

[18] Solvia (27 de marzo, 2018). 5 factores que influyen en el precio de una vivienda. https://www.solvia.es/magazine/5-factores-que-influyen-en-el-precio-de-una-vivienda/.

Glosario

Outlier	Vivienda que tiene valores atípicos.
préstamo hipotecario	préstamo para la compra de una Vivienda.
garantía hipotecaria	realización de un prestamo que no tiene como finalidad la compra de una vivienda, por el contrario una vivienda en propiedad se utiliza como aval de préstamo para otra actividad.
cobertura cartera viviendas	accion realizada por ciertos agente con la finalidad de minimizar las pérdidas ocasionadas por un movimiento desfavorable de los precios.
Cross Validation	técnica utilizada para evaluar resultados, de un análisis estadístico y garantizar que son independientes de la partición entre datos de entrenamiento y validación.